Oxford Internet Institute, University of Oxford

Assignment Cover Sheet

Candidate Number	1081481
Assignment	Introduction to Natural Language Processing for the Social Sciences
Term	Hilary Term 2024
Title/Question	Scientific Teams and Topic Diversity: A Semantic Analysis
Word Count	4996

By placing a tick in this box ☑ I hereby certify as follows:

(a) This thesis or coursework is entirely my own work, except where acknowledgments of other sources are given. I also confirm that this coursework has not been submitted, wholly or substantially, to another examination at this or any other University or educational institution;

(b) I have read and understood the Education Committee's information and guidance on academic good practice and plagiarism at `https://www.ox.ac.uk/students/academic/guidance/skills?wssl=1`.

(c) I agree that my work may be checked for plagiarism using Turnitin software and have read the Notice to Candidates which can be seen at: `http://www.admin.ox.ac.uk/proctors/turnitin2w.shtml`, and that I agree to my work being screened and used as explained in that Notice;

(d) I have clearly indicated (with appropriate references) the presence of all material I have paraphrased, quoted or used from other sources, including any diagrams, charts, tables or graphs.

(e) I have acknowledged appropriately any assistance I have received in addition to that provided by my [tutor/supervisor/adviser].

(f) I have not sought assistance from a professional agency;

(g) I understand that any false claims for this work will be reported to the Proctors and may be penalized in accordance with the University regulations.

Please remember:

- To attach a second relevant cover sheet if you have a disability such as dyslexia or dyspraxia. These are available from the Higher Degrees Office, but the Disability Advisory Service will be able to guide you.

Scientific Teams and Topic Diversity: A Semantic Analysis

1081481

Abstract

This study investigates whether the semantic diversity of paper titles, reflecting the range of topics covered in a research area, is associated with varying team sizes across scientific fields. Analyzing over one million publications using word embedding models, it compares the semantic diversity of paper titles with author counts. The findings reveal key insights into the relationship between collaboration and the breadth of knowledge produced within fields. This study is a valuable contribution to the science of science for a topic that has received recent attention: the role of team collaboration in knowledge production. It is of growing relevance as collaboration and communication technologies become more pervasive. Given this trend, understanding the influence of team sizes can help understand future developments in the diversity of scientific research.

1 Introduction and Literature Review

The dynamics of scientific knowledge production have been of increasing interest as data on scientific outputs becomes more readily available. The science of science has emerged as a relevant field to examine trends in scientific research, providing an increased understanding of knowledge dynamics that may inform more effective tools and policies for scientific progress (Fortunato et al., 2018). The dynamics of teams have received particular attention as collaboration becomes more common and is facilitated by information communication technologies. Innovation has also been central as stakeholders are interested in understanding innovation trends and maintaining the production of novel ideas.

Past literature has revealed a movement towards teams and an increase in team sizes across all scientific fields (Fortunato et al., 2018; Milojević, 2014). Even fields traditionally dominated by individual contributors, such as mathematics, have seen a rise in coauthorship and team dispersion since the 1990s, facilitated by decreasing collaboration costs (Adams et al., 2005). Notably, teams are not only growing in prominence but also tend to receive high citation counts and have greater scientific impact than individuals (Fortunato et al., 2018; Wuchty et al., 2007; Larivière et al., 2015). One possible explanation for this is that teams are able to produce more novel idea combinations (Fortunato et al., 2018). However, Wu et al. (2019) counter this, finding that smaller teams are more likely to disrupt a field with new ideas while larger teams tend to build on existing knowledge. Team diversity may also be a factor; for instance, teams with more gender diversity had more novel, high-impact ideas (Yang et al., 2022).

Paper titles have also received attention in research, demonstrating their ability to offer insights into field characteristics and determinants of scientific impact. For instance, studies have found that title style characteristics, including length and form, are highly related to their discipline, and that this has changed over time (Milojević, 2017; van Wesel et al., 2014; Yitzhaki, 1994). Several studies have also found a relationship between the number of authors and title characteristics; this includes a positive relationship with title length and a negative relationship with the use of punctuation such as colons and question marks (Lewison & Hartley, 2005; Hudson, 2016). Further, title semantic analyses have been conducted to identify differences in paper topic diversity. For instance, Hackett et al. (2021) compared topic diversity for papers coming from synthesis centers, and Jeon et al. (2023) examined titles and their embeddings to measure the novelty of publications. Titles have been used in many analyses as they are seen as a consolidated representation of the topic of a paper – more concise than abstracts but with more information than keywords (Milojević, 2017). Beyond communicating paper content, they reflect alignment to particular norms in a field or community and represent an academic identity (Milojević, 2017).

This study extends these analyses by examining the relationship between team size and the semantic diversity of paper titles. Drawing from the prior literature on team size and scientific novelty and employing large-scale scientific paper data, it asks: Are changes in team size related to differences in the diversity of paper topics? To analyze the semantic diversity of paper titles, I employ word embeddings from a vector space model, leveraging fastText for its ability to represent out-of-vocabulary words. The input variable is team size, defined by the number of authors credited on a paper. The output variable is semantic diversity of titles, which is operationalized to represent paper topic diversity in a field.

This study contributes to bodies of literature on title analysis, scientific collaboration, and research topic diversity. It provides a novel analysis that operationalizes scientific paper titles to contribute to knowledge on team dynamics in relation to field diversity. If the results show that larger team sizes have higher diversity of titles, this may suggest that larger teams can contribute to more diverse knowledge. If the opposite is true, it may be that larger teams stifle diversity in research and converge towards a narrower scope within a field. This body of knowledge can inform future decisions to foster collaboration or independent work with the aim of advancing scientific progress and innovation.

2 Methods

2.1 Data

This paper uses a dataset of academic papers from ArXiv, which contains attributes of papers from eight fields. For the purposes of this analysis, I downloaded the arXiv metadata obtained from Kaggle (arXiv.org submitters, 2024) and included the titles of scientific papers, the list of authors, the category labels, and the initial publication year. I used the initial publication year as opposed to subsequent updates in order to represent the date at which the topic of the paper was initially presented. Each paper in the dataset has a primary category, which is the first category listed, and may have additional category labels (Clement et al., 2019). In this paper, such categories are referred to as subcategories, while the main category is the indicator at the beginning of the subcategory name. For instance, a paper's subcategory would be math.it (Information Theory) and its main field or category would be math. The titles are the principle unit of data used to extract information about the paper topics for the reasons mentioned above.

To prepare the titles for analysis, I converted them all to lowercase, removed any punctuation with the exception of hyphenated words, and removed any abbreviations in parentheses as suggested by Jeon et al. (2023). Additionally, I removed stop words as the primary focus of the analysis concerns the topic of discussion. Unlike a measure such as sentiment in which stop words could be meaningful, removing common words and focusing on substantive words provides a better representation of the paper topic (Yitzhaki, 1994). I chose to focus on three major fields: mathematics, computer science, and physics. These fields contain a large number of data points and subcategories for the analysis and provide an interesting comparison as mathematics has traditionally been dominated by sole authors, physics has become known for extremely large teams, and computer science is a rapidly developing field with increasing collaboration trends.

2.2 Descriptive Statistics

Figures 1-3 show the number of papers in the dataset by team size for different categories and years. In total, there are 1,142,849 papers in physics, math, and computer science. There are 30 subcategories in math, 40 in computer science, and 22 in physics. The number of papers substantially increased in all categories at least through 2020. As expected, there are notable differences in the authorship between fields. Mathematics has had a more substantial and consistent proportion of solo authors, while teams make up the majority in computer science and physics. As the data prior to 2000 is more volatile and sparse, the present analysis focuses on trends from the year 2000 to 2023. This leaves 1,134,092 papers in the dataset for the fields of mathematics, computer science, and physics. There are 478,301 papers from mathematics, 484,511 from computer science and 171,280 from physics.

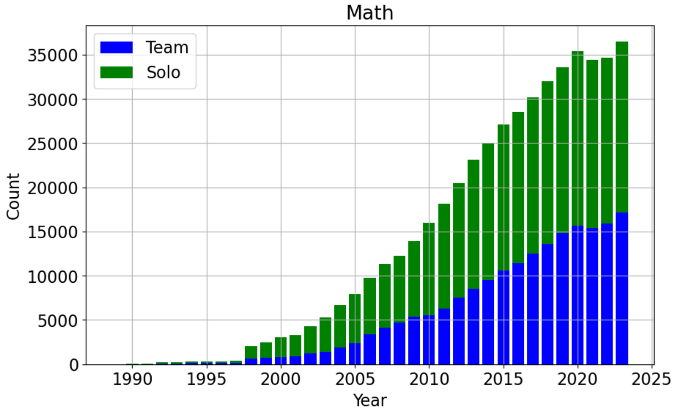

Figure 1: Papers by year and team type: mathematics

Figure 4 shows the distribution of papers by team size in the last five years. Most mathematics papers have less than five authors and most often have one, while in computer science and physics many have 5 to 10 authors. There are also substantial

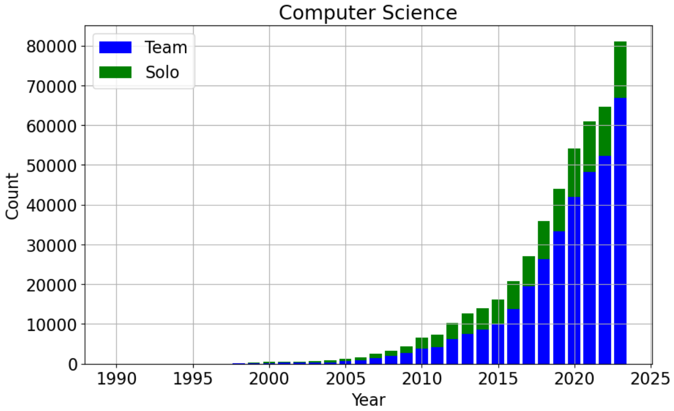

Figure 2: Papers by year and team type: computer science

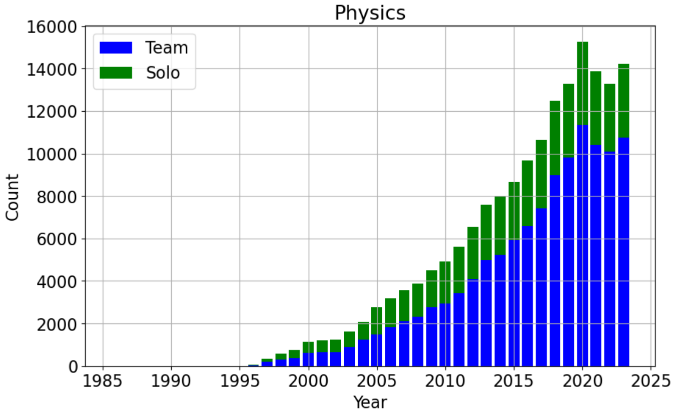

Figure 3: Papers by year and team type: physics

outliers not shown in the histogram for visualization purposes. Table 2.2 shows the descriptive statistics of all team size data per field. Some papers involve mass collaborations with hundreds of authors; physics is particularly notable for this with the largest team size at 1,324 authors. Computer science also sees some large teams, with the largest having 941 authors. Nonetheless, the distribution is largely concentrated below 5 for all fields, and below 2 for mathematics.

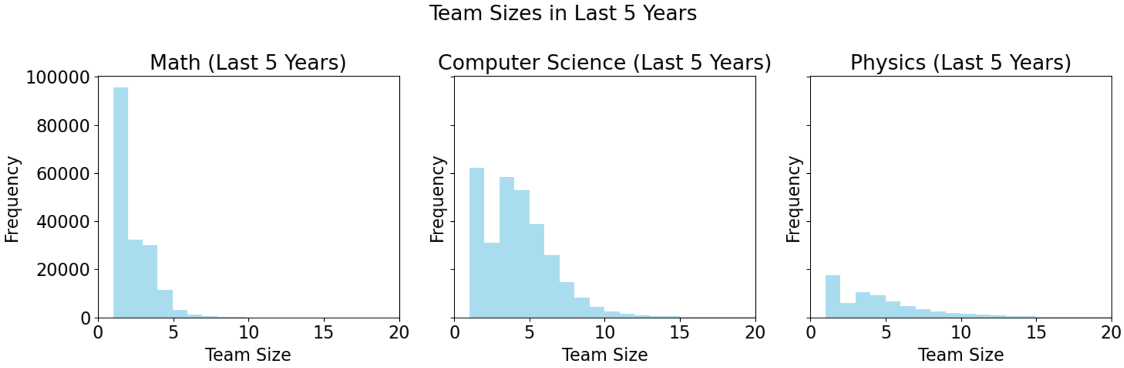

Figure 4: Distribution of Team Sizes in Last Five Years

Statistic	Mathematics	Computer Science	Physics
Count	478,301	484,511	171,280
Mean	1.73	3.53	4.90
Std. Dev.	1.12	3.40	16.66
Min	1	1	1
25%	1	1	1
50% (Median)	1	3	3
75%	2	5	5
Max	82	941	1,324

Table 1: Descriptive statistics of team sizes by field

2.3 Computational Architecture

2.3.1 Word Embeddings

In order to create a representation of the semantic similarity of papers, I utilized word embeddings to represent the location of words and subsequently titles in relation to each other. Word embeddings provide representations of words in a vector space. Word embedding systems use distributed representations of dense vectors, which typically have 50-1000 dimensions, are shorter and perform better than sparse vectors (Jurafsky & Martin, 2024). They are better for representing synonyms and, with fewer dimensions, the models require learning less weights which can reduce overfitting (Jurafsky & Martin, 2024). Vector semantics represent the meaning of words and stem from the idea that the meaning of the word can be derived from the words surrounding it (Jurafsky & Martin, 2024). Thus, words representing similar topics will be located closer to each other in the vector space. The difference between them can be measured by the cosine angle of the two vectors – the most common metric for similarity (Jurafsky & Martin, 2024). The more similar they are, the closer the cosine similarity will be to one. To obtain a representation of an entire title, I took the average of the vectors of the words in the title to get the title vector. This is a common practice to represent documents or titles containing multiple words (Jeon et al., 2023). With a vector for each title, I was then able to derive its semantic similarity by calculating its cosine similarity, or 1 minus the cosine distance, to another vector.

2.3.2 fastText

I employed a fastText model to obtain the vectors for each kept word in the titles. The fastText word-embedding model was developed by the Facebook AI Research Lab (Bojanowski et al., 2017). fastText builds upon Word2Vec, developed by Mikolov et al. (2013), which proposed skip-gram and Continuous Bag-of-Words (CBOW) models to generate vector representations for words. The CBOW model leverages a target word's context, or the words surrounding it, to train a log-linear classifier and predict a target word (Mikolov et al., 2013). CBOW then aims to maximize the log-likelihood of the probability of words considering their surrounding context (Mikolov et al., 2017). The skip-gram model does the opposite, using a target word to predict the surrounding context. (Mikolov et al., 2013; Garcia & Gomez-Perez, 2018; Rong, 2016). This development was highly valued for its speed, as it randomly skips over words, uses only one hidden layer, and uses small text windows.

FastText can use either of the skip-gram or CBOW architectures from Word2Vec, but offers an advantage as Word2Vec cannot address out-of-vocabulary words that do not exist in the training data (Jeon et al., 2023). Meanwhile, fastText is trained on character n-grams, which solves the sparsity issue as it can derive a vector for an unknown word by using a bag of its character n-grams, or sub-words (Jeon et al., 2023). Thus, the word vectors represent a word by averaging over its character

n-grams (Bojanowski et al., 2017). This is a relevant issue to address in a dataset of highly scientific words where new concepts or terms may often be introduced.

2.3.3 Model Selection and Training

The most recently provided model from fastText contains pre-trained vectors trained on Common Crawl and Wikipedia. Common Crawl is a repository of publicly available web crawl data (*Common Crawl*, n.d.) and, while it can introduce noisier data as it draws from the broader web, it can also provide better coverage for models (Grave et al., 2018). The fastText model used a CBOW (Continuous Bag of Words) architecture in 300 dimensions. They used character n-grams with a length of 5, a window size of 5 and 10 negatives (Grave et al., 2018). This model added position-dependent weighting, which improves its sensitivity to the word order in the context window (Grave et al., 2018). It used the model proposed by Mnih and Kavukcuoglu (2013) with position weights, where word vectors are multiplied by a position dependent vector before taking the sum (Grave et al., 2018; Mnih & Kavukcuoglu, 2013). This CBOW model with position weights, using 10 negative examples and trained for 10 epochs had the best performance on word analogy tasks. Other tested models included the fastText skip-gram model with default parameters and models trained only on Wikipedia (Grave et al., 2018). Using the position weights gave the largest improvement in model accuracy of all modifications tested.

As the pre-trained fastText model was trained on Common Crawl and Wikipedia, the embeddings may not be particularly suited for scientific papers and they may appear in a similar space in the semantic cloud. Therefore, I further trained the fastText model on the titles from the ArXiv dataset. This follows the design by Jeon et al. (2023) with the objective to improve the representation of the semantic relationships between the words in paper titles. In the further training of the model, I maintained most pre-trained model hyperparameters but changed the window size to 15 to capture the full context of the majority of titles, given that most titles have a word length under 15 (see Figure 5). Similar to Jeon et al. (2023), the model was trained over 200 epochs to ensure a thorough learning of the semantic relationships in the training text. Additionally, although it increases training time, Grave et al. (2018) found that more epochs resulted in significantly higher accuracy for the models. To select the best model, I trained additional models using different window sizes, epochs, and baseline models to compare the model improvement. To prepare the titles for the model training, I divided the cleaned titles into separate words. The word-level tokenization is used as the unit for model training, which is the default configuration for the pre-trained model and follows the architecture used by Jeon et al. (2023).

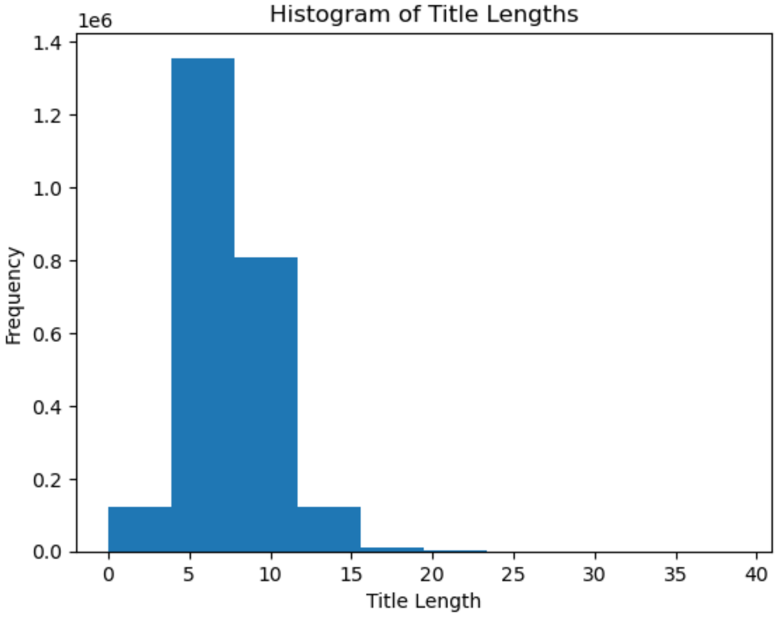

Figure 5: Distribution of number of words in titles

2.4 Measurement

The measure of uniqueness of a paper is represented by the semantic similarity of that paper to the centroid of its subfield. The higher its semantic similarity, the less unique it is in the context of the subfield. The subfield centroid is calculated by taking the average of the vectors of all titles in that subfield. Each paper was compared to a centroid of its subfield consisting of all the papers existing in that subfield up to that paper's creation year.

To examine the relationship between team sizes and paper uniqueness, I use a similarity ratio (SR) of teams to individuals. This measure draws on the metric designed by Wuchty et al. (2007) to evaluate citation impact for teams vs. individuals. Here, the relative measure takes the mean semantic similarity for team-authored work divided by the mean semantic similarity of individual work. Therefore, a ratio greater than 1 would show that teams have higher semantic similarity to the average paper in their subfield; in other words, teams would have less unique titles relative to individuals.

3 Results

3.1 Model Validation

To validate the quality of the further trained fastText model, I calculated the average pairwise similarity of the words within each title, following the design by Jeon et al. (2023) to validate the quality of further trained fastText models for scientific papers. The further trained model should have a higher similarity between the words in a title than the original to indicate an improvement in the semantic relationships of the article title words. Jeon et al. (2023) represent the average pairwise similarity metric as follows:

$$average\ pairwise\ similarity\ (p) = \frac{\sum_{k_i \in p} \sum_{k_j \in p, j>i} \cos(k_i, k_j)}{N_{pk}}$$

Here, p is a paper, k_i and k_j represent two words in the paper's title, and $p, cos(k_i, k_j)$ is the cosine similarity between k_i and k_j. N_{pk} represents the total words in the title for paper p (Jeon et al., 2023). In other words, it takes the average cosine similarity between all word pairs in a title, for all titles.

I first tested pre-trained fastText English models trained on only Wikipedia data as well as on Common Crawl (CC). The Wikipedia model had 10 unique words not in the model vocabulary. The model trained on Wikipedia and Common Crawl only had one unique word not present. In further training, I tested several window sizes and trained over 100 and 200 epochs. Table 2 shows the average similarity found for the baselines and top-performing further trained models. The Wikipedia model was trained for 100 epochs as it had a notably slower training time. Increasing the window size from 15 to the window size used by Jeon et al. (2023) (2^{30}) to include all word pairs did not result in increased pairwise similarity. The Common Crawl model initially had the lowest similarity between words in a title, but with further training it saw a substantial increase. I found the most improvement and the highest similarity on the Common Crawl + Wikipedia model with a window size of 15, increasing from 0.19 to 0.54. These baseline and post-training similarities are in line with the similarities found by Jeon et al. (2023) for their fastText model and academic paper dataset. Therefore, I opted to use the further-trained model using Common Crawl and Wikipedia data with a window size of 15 to obtain embeddings for the remaining analysis.

Model	Average Similarity
Baseline Wiki	0.271910
Baseline CC and Wiki	0.186672
Further trained Wiki	0.322880
Further trained CC and Wiki, window 15	0.539340
Further trained CC and Wiki, extreme window	0.507860

Table 2: Average similarity of words in a paper title for fastText models

To test the model's embeddings, I also examined the distance between each subfield to validate the expected distances. I used the 2023 centroid of each subfield, which was calculated by averaging the vectors of all the paper titles in that subfield through 2023. Figure 2 shows the cosine distance between each subfield ordered by major category. As expected, the dark areas show that the subfields within each field are closer to each other than to the other two fields. Physics and mathematics also appear generally closer than mathematics and computer science, though there are some dark streaks for particular subfields in mathematics and computer science that are closely related. Additionally, I performed the same analysis for the baseline CC + Wikipedia model to further test the model quality. Figure A.1 shows larger distances between similar subfields and less distinctiveness between the major fields. This further demonstrates the improvement of the further trained model for the article title embeddings.

3.2 Relative Team Semantic Similarity Over Time

An analysis of team sizes in Figure 7 shows an evident increase in team sizes over time. The mathematics field remained the most consistent with the average team size staying below two. Meanwhile, computer science particularly increased after 2010, with the average doubling from around 2 in 2000 to over 4 in 2020. Physics saw the largest increase with the average team size reaching 6 after 2020.

The semantic similarity measures showed a less consistent pattern. Figures A.5-A.10 show the distribution of paper semantic diversity per field for the first five and last five years of the analysis. Computer science has the largest spread and mathematics the smallest, though most papers have a high similarity to their subfield average. Papers also concentrated slightly closer to their subfield average over time. Figure 8 shows the ratio of teams to individuals for the average semantic similarity in each major field. Mathematics had the most consistent ratio, remaining slightly above one

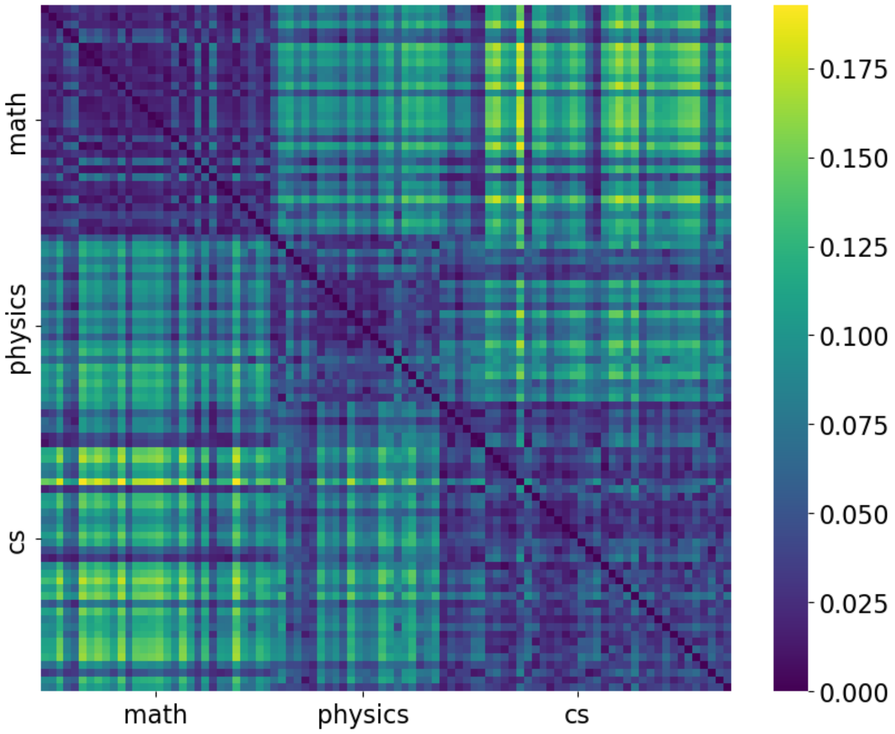

Figure 6: Cosine distance between subfield centroids: further trained fastText model

with little variation over time. The computer science field saw more variation and notably increased after 2015 when team sizes also noticeably increased, though it also did not deviate far from one. In physics, the highest ratio from team to individual similarity was seen, though it remained lower than 1.03. As team sizes increased, the similarity ratio had an overall decline beginning in 2010. Generally, teams almost always had higher semantic similarity than individuals, but the magnitude was small.

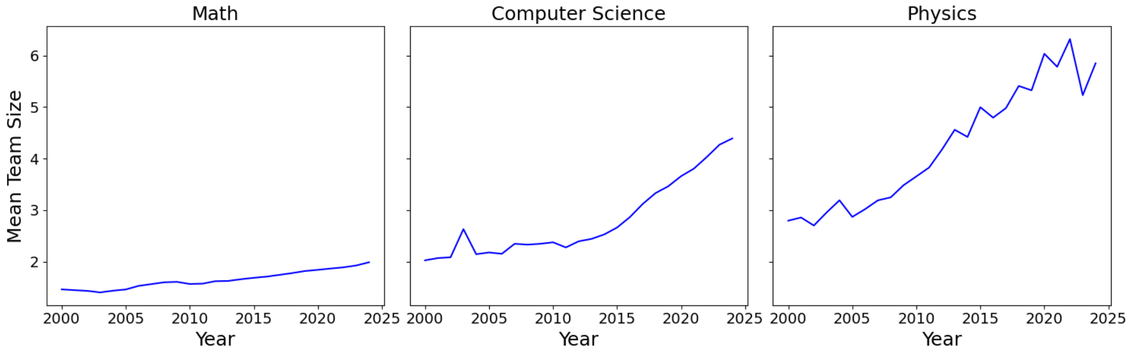

Figure 7: Mean team sizes over time by field

Figure 9 shows the average similarity ratio for all subcategories, where each point is the ratio for a given subcategory and year. It shows that over time, most sub-

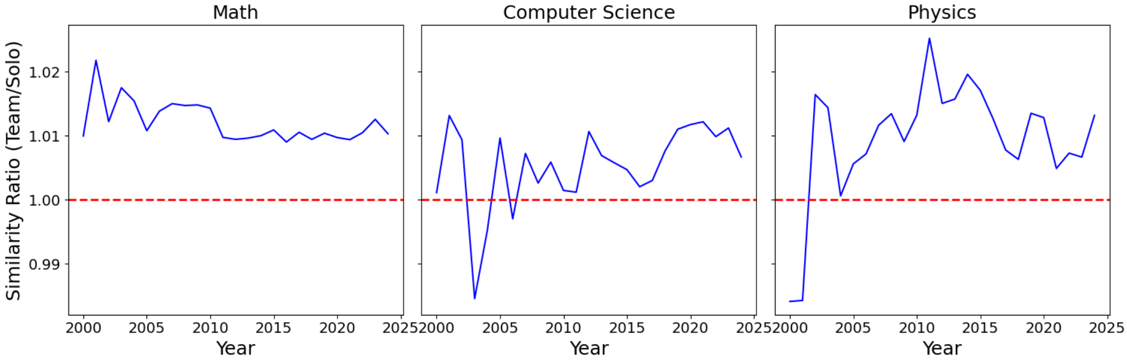

Figure 8: Arithmetic average of semantic similarity ratio by year

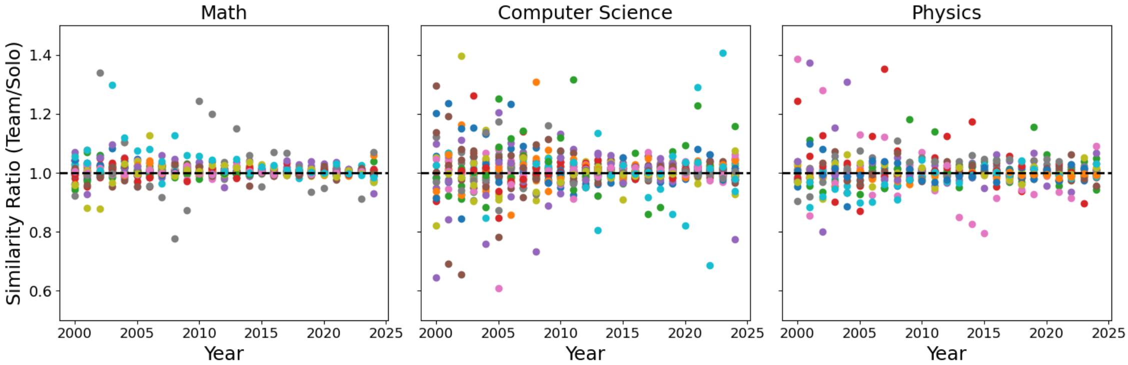

Figure 9: Semantic similarlity ratio by subfield: Mean semantic similarity of teams for each subcategory divided by mean semantic similarity of solo authors. Each point represents a given subcategory and year.

categories converge towards less variation in semantic similarity between teams and individuals. The mathematics field saw the least variation, with most subfield ratios remaining close to one. In computer science, however, there were a number of subfields that had a notable difference between teams and individuals. However, the variation was seen on both sides; in some subfields teams were up to 1.4 times more similar to the average than individuals, and in others teams were less than 0.8 times as similar. In physics many of the subfields also converged towards one over time. Overall, there are more subfields with teams having less unique titles than solo authors, but not by a substantial amount.

I also calculated similarity ratios to individuals for only small teams (2-5 authors), medium teams (6-10 authors), and large teams (11+ authors), which are shown in figures 10-12. The ratios by subfield are presented in A.2-A.4. As the team size increased, the variation in the ratios increased, while the small team ratios remained close to one. Notably for large teams, the ratio dropped below 1 in recent years for all three fields.

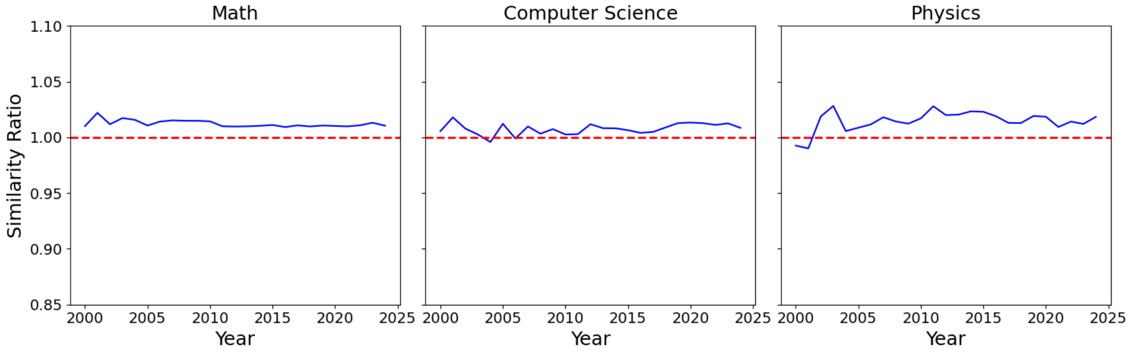

Figure 10: Average semantic similarity ratio of small teams (2-5 authors) to individuals for all papers by year

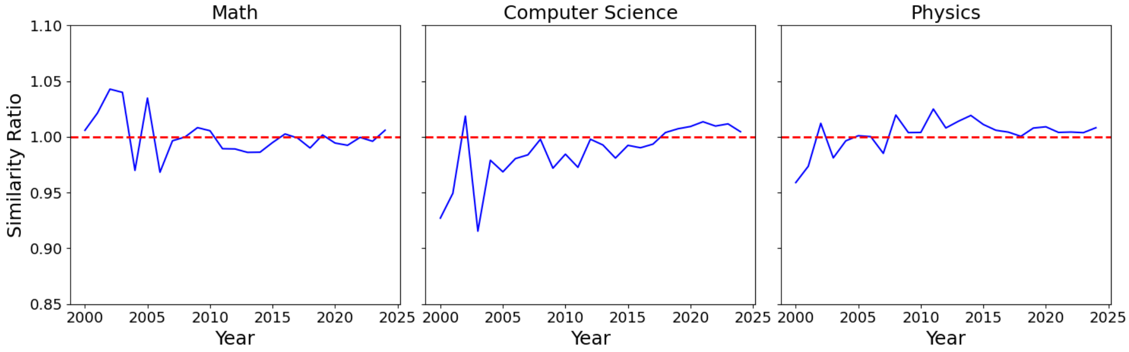

Figure 11: Average semantic similarity ratio of medium teams (6-10 authors) to individuals for all papers by year

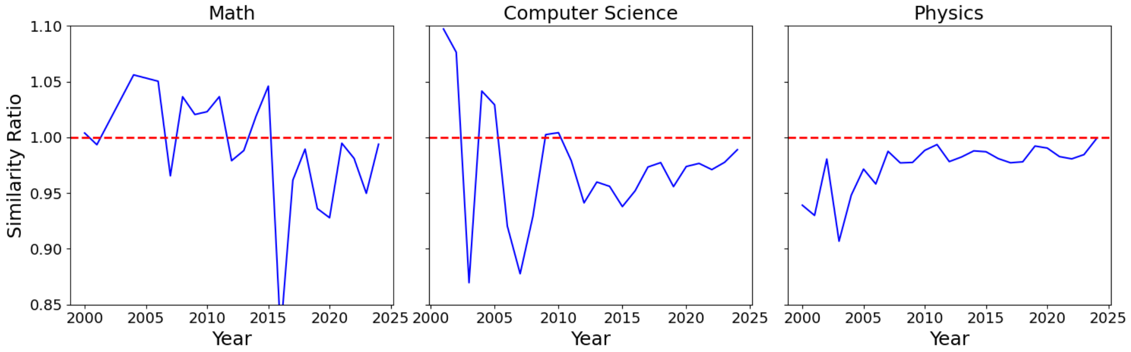

Figure 12: Average semantic similarity ratio of large teams (11+ authors) to individuals for all papers by year

3.3 Team Size vs. Semantic Similarity Correlation Analysis

I also examined the relationship between paper cosine similarity and team size on a paper level for each subfield in the last five years. I calculated the Spearman correlation between the cosine similarities and team sizes given the non-normal data distribution. Table 3 contains the correlations for each subfield. The correlations are generally low and insignificant. The mathematics field has the least subfields with significant correlations, though those that are significant are mostly negative with the exception of General Mathematics (math.GM). Computer science has more significant correlations, though they vary in direction. Physics has the most subfields with significant correlations and they are nearly all negative, suggesting that there is a negative relationship between team size and semantic similarity of titles within physics subfields. In other words, larger teams produce more unique titles, though the magnitude of the correlation is still small.

4 Conclusions and Limitations

The findings from over one million data points suggest that teams, mostly composed of 2 to 10 authors, do not necessarily lead to more diverse or unique research topics as measured by title semantics. The ratios of overall team to individual semantic similarity indicate there is not a substantial difference between teams and individuals. The varied and small correlations on a paper level between team size and semantic similarity further suggest a weak relationship. Thus, general collaboration does not necessarily lead to more diverse scientific papers. Likewise, there is limited evidence that teams stifle paper uniqueness. This observation contrasts propositions

Math	Correlation	CS	Correlation	Physics	Correlation
math.CO	0.01	cs.NE	0.03	physics.gen-ph	0.01
math.CA	0.01	cs.DS	0.01	physics.optics	-0.12***
math.NT	-0.00	cs.CE	-0.09***	physics.ed-ph	0.07*
math.PR	0.00	cs.IT	0.01	physics.pop-ph	-0.00
math.NA	-0.02*	cs.CC	0.03	physics.soc-ph	-0.03*
math.RA	0.02	cs.DM	0.02	physics.data-an	-0.00
math.OA	-0.01	cs.CR	0.04***	physics.plasm-ph	-0.09***
math.QA	0.02	cs.NI	-0.02	physics.bio-ph	-0.05*
math.DG	0.01	cs.PF	0.03	physics.flu-dyn	-0.07***
math.FA	0.01	cs.LG	-0.03***	physics.comp-ph	-0.07***
math.AG	0.01	cs.CY	0.03*	physics.atom-ph	-0.05*
math.DS	-0.02	cs.CG	0.01	physics.chem-ph	-0.14***
math.GR	0.02	cs.CV	-0.05***	physics.geo-ph	-0.06*
math.AC	-0.00	cs.SE	-0.02	physics.class-ph	0.07*
math.SG	0.02	cs.OH	0.05	physics.atm-clus	-0.05
math.GT	0.00	cs.PL	0.04	physics.acc-ph	-0.09***
math.CV	0.02	cs.AI	0.01	physics.hist-ph	-0.06
math.AP	-0.03***	cs.IR	0.08***	physics.ins-det	-0.03*
math.RT	0.01	cs.GT	0.02	physics.space-ph	0.07*
math.MG	0.00	cs.LO	0.00	physics.med-ph	-0.08***
math.ST	-0.00	cs.SC	0.04	physics.ao-ph	-0.08***
math.AT	0.02	cs.DC	0.00	physics.app-ph	-0.08***
math.OC	-0.03***	cs.CL	0.02***		
math.LO	0.04*	cs.HC	0.02		
math.GM	0.05*	cs.AR	-0.00		
math.KT	0.04	cs.DL	0.05		
math.SP	-0.00	cs.MS	0.05		
math.HO	-0.08***	cs.RO	-0.05***		
math.CT	-0.00	cs.DB	-0.00		
math.GN	0.02	cs.GL	0.05		
		cs.MA	0.06*		
		cs.MM	-0.00		
		cs.OS	0.10		
		cs.NA	-0.06		
		cs.SD	-0.01		
		cs.GR	0.05		
		cs.FL	-0.02		
		cs.SI	-0.05***		
		cs.SY	0.05		
		cs.ET	0.07*		

Table 3: Cosine similarity and team size correlation for papers in each subfield in the last five years (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$)

that teams garner a wider range of ideas, or conversely, that individuals are more likely to produce fringe work.

However, the further analysis by different team sizes suggests that the pattern may change for large teams. While the similarity ratios were generally above one for small teams, they fell below one for large teams particularly in recent years. The overall decrease in the similarity ratio for physics may also reflect this as large teams became more prevalent. Thus, large teams may differ from the overall trend in that they produce more unique paper titles than individuals. This highlights that team dynamics may have a more complex relationship with innovation. It is also in line with prior findings that large teams diverge from general team patterns (Lewison & Hartley, 2005).

Past literature has shown that teams see different research outcomes, such as receiving higher citations (Fortunato et al., 2018; Wuchty et al., 2007; Larivière et al., 2015; van Wesel et al., 2014). However, citation impact does not necessarily relate to measures of a title's semantic diversity. The analysis also contrasts the finding that smaller teams have more disruptive ideas (Wu et al., 2019). Nonetheless, disruptiveness in the study by Wu et al. (2019) represented novelty through a citation-based metric and differs from the measure of semantic similarity in titles. Thus, it is possible that less authors have more disruptive ideas but do not necessarily have more unique titles. Other literature has found ambiguous relationships between title characteristics and team sizes (Yitzhaki, 1994), which support the finding that titles may not differ substantially between teams and individuals.

Methodologically, this paper presented a viable application of word embeddings to measure title uniqueness based on its semantic similarity to the average paper in its subfield. This provides an accessible approach to analyze paper diversity as it employs title data, which is more efficient and requires less preprocessing than larger bodies of text (Jeon et al., 2023). It is feasible as titles are a representative summary of the ideas presented in a paper. Further, the trained model's suitability for the scientific context was validated as the embeddings showed expected distances between title words and subfields. This method can be useful in numerous research applications aiming to measure topic diversity.

Though the findings provided valuable insights, there are limitations to consider in their interpretation. First, the analysis can make no causal claims. Nonetheless, the findings suggest that the relationship between team collaboration and title semantic diversity is limited. Though beyond the scope of this paper, future research could employ more advanced statistical analyses to further explore the relationship between team size and topic diversity, understand potential confounding factors, and uncover the direction of the relationship, if any. It should also be noted that the study only employs one measure of paper uniqueness using semantic similarity of titles. Jeon et al. (2023) argued that most titles aim to reflect a paper's core ideas and optimize platform searches. Nonetheless, titles may also be strategic to attract readers and use language such as metaphors (Jamali & Nikzad, 2011). Further, they may overlook deeper nuances that other paper sections could reveal (Jeon et al., 2023), and the analysis showed limited diversity between most titles. Thus, future research could analyze larger volumes of text from papers to gain a more thorough representation of paper content. It could also compare other measures of paper novelty or diversity to gauge alignment with the title measure.

A few limitations were also present due to the computational resources available. First, the word embeddings were static, and the same model was used to obtain word embeddings for all years. Future analyses could train separate models for different years to have a more precise word embedding at a given point in time. Additionally, dynamic contextual embeddings like BERT representations, which provide different vectors for different contexts, could be used (Jurafsky & Martin, 2024). For instance, Sentence-BERT could be applicable to obtain sentence embeddings for the titles (Reimers & Gurevych, 2019). Additionally, though several hyperparameters were tested in training the fastText models, a more thorough approach using grid or random search could be useful to find the optimal hyperparameters. As training each fastText model took a substantial amount of time and computational resources were limited, these extensions were not possible in this analysis.

Despite the limitations, these findings contribute to the ongoing discourse on how team size influences research diversity, an area of focus as collaborative trends advance. They provided insight on the complex relationship between team collaboration and paper uniqueness, highlighting that although smaller teams did not substantially differ from individuals, the recent uptick in large teams could alter patterns of knowledge production. Further research could also explore other dimensions of team diversity, such as disciplinary and cultural backgrounds, to provide a more comprehensive understanding of how these factors impact scientific novelty. This adds to the academic literature on the science of science to understand current and future trends in the diversity of knowledge outputs. Additionally, knowledge about collaboration impacts is essential for decisions to structure teams that facilitate innovation while optimizing resource allocation. This study therefore provides a valuable contribution using natural language processing methods and large-scale data to advance our understanding of team dynamics in knowledge production.

References

Adams, J. D., Black, G. C., Clemmons, J. R., & Stephan, P. E. (2005, April). Scientific teams and institutional collaborations: Evidence from U.S. universities, 1981–1999. *University-based Technology Initiatives*, *34*(3), 259–285. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0048733305000132` doi: 10.1016/j.respol.2005.01.014

arXiv.org submitters. (2024). *arXiv Dataset.* Kaggle. Retrieved from `https://www.kaggle.com/dsv/7548853` doi: 10.34740/KAGGLE/DSV/7548853

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146.

Clement, C. B., Bierbaum, M., O'Keeffe, K. P., & Alemi, A. A. (2019). *On the Use of ArXiv as a Dataset.* (_eprint: 1905.00075)

Common Crawl. (n.d.). Retrieved 2024-05-03, from `https://commoncrawl.org/`

Fortunato, S., Bergstrom, C. T., Börner, K., Evans, J. A., Helbing, D., Milojević, S., . . . Barabási, A.-L. (2018, March). Science of science. *Science*, *359*(6379), eaao0185. Retrieved 2024-03-22, from `https://doi.org/10.1126/science.aao0185` (Publisher: American Association for the Advancement of Science) doi: 10.1126/science.aao0185

Garcia, A., & Gomez-Perez, J. M. (2018). *Not just about size - A Study on the Role of Distributed Word Representations in the Analysis of Scientific Publications.* (_eprint: 1804.01772)

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. (2018). Learning Word Vectors for 157 Languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018).*

Hackett, E. J., Leahey, E., Parker, J. N., Rafols, I., Hampton, S. E., Corte, U., . . . Vision, T. J. (2021, January). Do synthesis centers synthesize? A semantic analysis of topical diversity in research. *Research Policy*, *50*(1), 104069. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0048733320301475` doi: 10.1016/j.respol.2020.104069

Hudson, J. (2016, November). An analysis of the titles of papers submitted to the UK REF in 2014: authors, disciplines, and stylistic details. *Scientometrics*, *109*(2), 871–889. Retrieved from `https://doi.org/10.1007/s11192-016-2081-4` doi: 10.1007/s11192-016-2081-4

Jamali, H. R., & Nikzad, M. (2011, August). Article title type and its relation with the number of downloads and citations. *Scientometrics*, *88*(2), 653–661. Retrieved from `https://doi.org/10.1007/s11192-011-0412-z` doi: 10.1007/s11192-011-0412-z

Jeon, D., Lee, J., Ahn, J. M., & Lee, C. (2023, November). Measuring the novelty of scientific publications: A fastText and local outlier factor approach. *Journal of Informetrics*, *17*(4), 101450. Retrieved from `https://www.sciencedirect.com/science/article/pii/S1751157723000755` doi: 10.1016/j.joi.2023.101450

Jurafsky, D., & Martin, J. H. (2024). *Speech and Language Processing* (3rd ed.). Retrieved from `https://web.stanford.edu/~jurafsky/slp3/`

Larivière, V., Gingras, Y., Sugimoto, C. R., & Tsou, A. (2015, July). Team size matters: Collaboration and scientific impact since 1900. *Journal of the Association for Information Science and Technology*, *66*(7), 1323–1332. Retrieved 2024-03-22, from `https://doi.org/10.1002/asi.23266` (Publisher: John Wiley & Sons, Ltd) doi: 10.1002/asi.23266

Lewison, G., & Hartley, J. (2005, April). What's in a title? Numbers of words and the presence of colons. *Scientometrics*, *63*(2), 341–356. Retrieved from `https://doi.org/10.1007/s11192-005-0216-0` doi: 10.1007/s11192-005-0216-0

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space.* (_eprint: 1301.3781)

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., & Joulin, A. (2017). *Advances in Pre-Training Distributed Word Representations.* (_eprint: 1712.09405)

Milojević, S. (2014, March). Principles of scientific research team formation and evolution. *Proceedings of the National Academy of Sciences*, *111*(11), 3984–3989. Retrieved 2024-03-22, from `https://doi.org/10.1073/pnas.1309723111` (Publisher: Proceedings of the National Academy of Sciences) doi: 10.1073/pnas.1309723111

Milojević, S. (2017). The Length and Semantic Structure of Article Titles—Evolving Disciplinary Practices and Correlations with Impact. *Frontiers in Research Metrics and Analytics*, *2*. Retrieved from `https://www.frontiersin.org/articles/10.3389/frma.2017.00002`

Mnih, A., & Kavukcuoglu, K. (2013). Learning word embeddings efficiently with noise-contrastive estimation. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2* (pp. 2265–2273). Red Hook, NY, USA: Curran Associates Inc. (event-place: Lake Tahoe, Nevada)

Reimers, N., & Gurevych, I. (2019). *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks.* (_eprint: 1908.10084)

Rong, X. (2016). *word2vec Parameter Learning Explained.* (_eprint: 1411.2738)

van Wesel, M., Wyatt, S., & ten Haaf, J. (2014, March). What a difference a colon makes: how superficial factors influence subsequent citation. *Scientometrics*, *98*(3), 1601–1615. Retrieved from `https://doi.org/10.1007/s11192-013-1154-x` doi: 10.1007/s11192-013-1154-x

Wu, L., Wang, D., & Evans, J. A. (2019, February). Large teams develop and small teams disrupt science and technology. *Nature*, *566*(7744), 378–382. Retrieved from `https://doi.org/10.1038/s41586-019-0941-9` doi: 10.1038/s41586-019-0941-9

Wuchty, S., Jones, B. F., & Uzzi, B. (2007, May). The Increasing Dominance of Teams in Production of Knowledge. *Science*, *316*(5827), 1036–1039. Retrieved 2024-04-30, from `https://doi.org/10.1126/science.1136099` (Publisher: American Association for the Advancement of Science) doi: 10.1126/science.1136099

Yang, Y., Tian, T. Y., Woodruff, T. K., Jones, B. F., & Uzzi, B. (2022, September). Gender-diverse teams produce more novel and higher-impact scientific ideas. *Proceedings of the National Academy of Sciences*, *119*(36), e2200841119. Retrieved 2024-03-22, from https://doi.org/10.1073/pnas.2200841119 (Publisher: Proceedings of the National Academy of Sciences) doi: 10.1073/pnas.2200841119

Yitzhaki, M. (1994, May). Relation of title length of journal articles to number of authors. *Scientometrics*, *30*(1), 321–332. Retrieved from https://doi.org/10.1007/BF02017231 doi: 10.1007/BF02017231

5 Appendix

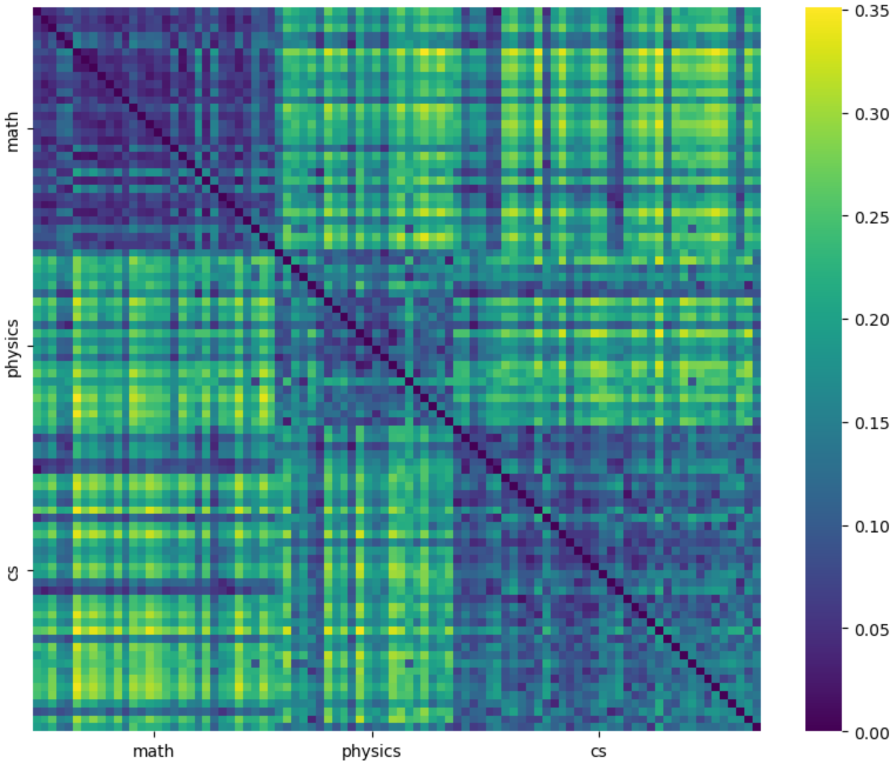

Figure A.1: Cosine distances between subfield centroids: baseline fastText model

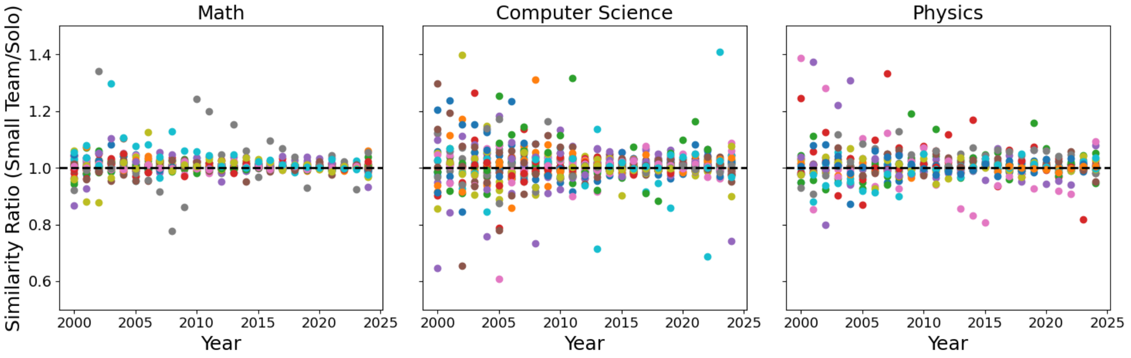

Figure A.2: Semantic similarlity ratio of small teams (2-5 authors) to individuals by subfield. Each point represents a given subcategory and year.

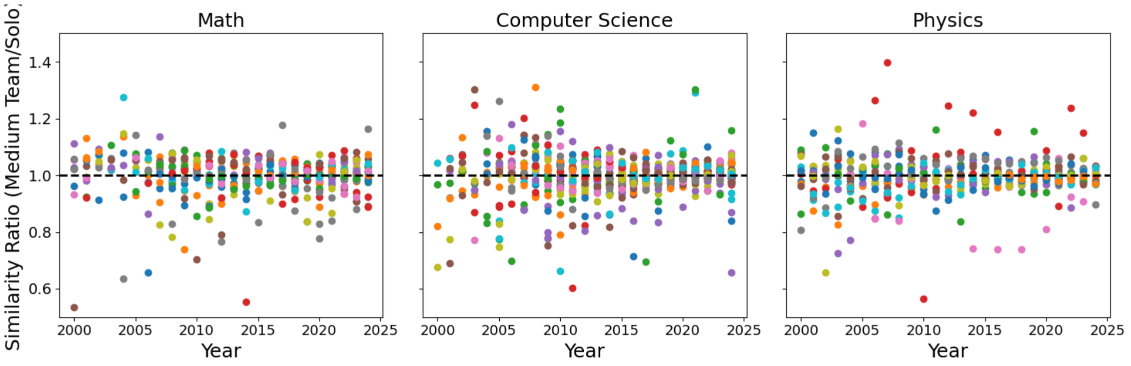

Figure A.3: Semantic similarlity ratio of medium teams (6-10 authors) to individuals by subfield. Each point represents a given subcategory and year.

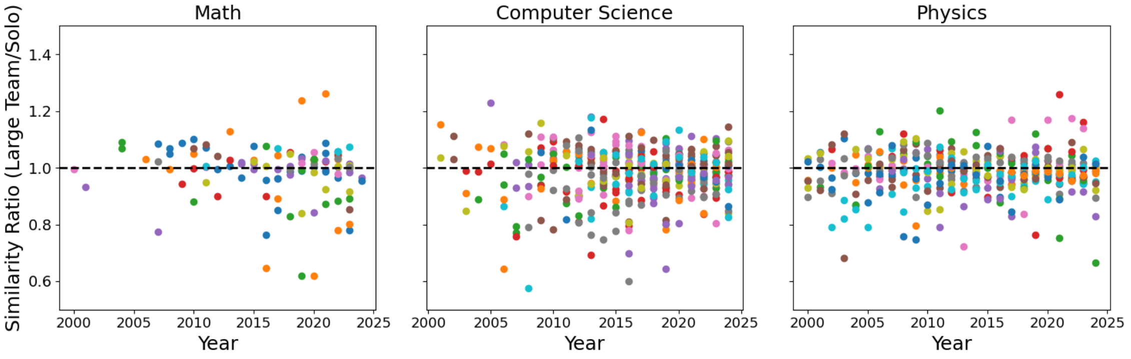

Figure A.4: Semantic similarlity ratio of large teams (11+ authors) to individuals by subfield. Each point represents a given subcategory and year.

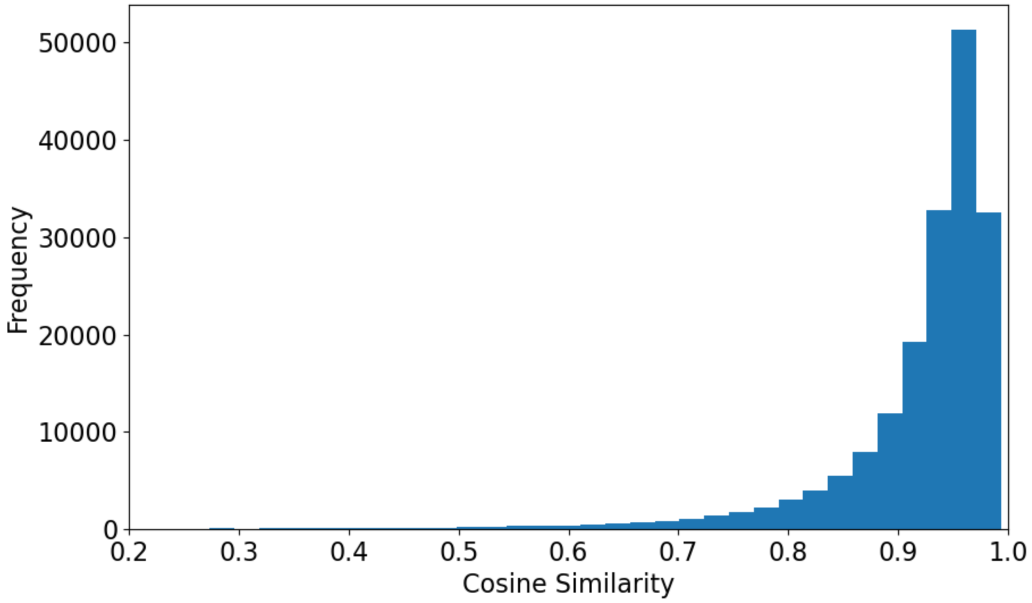

Figure A.5: Distribution of semantic similarities of all mathematic papers, last five years

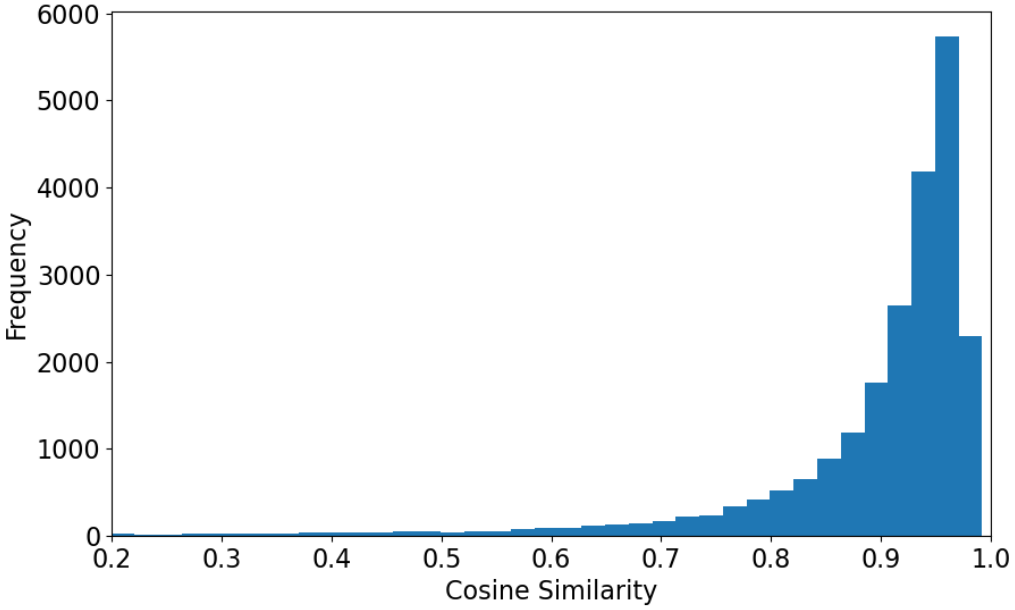

Figure A.6: Distribution of semantic similarities of all mathematic papers, 2000-2004

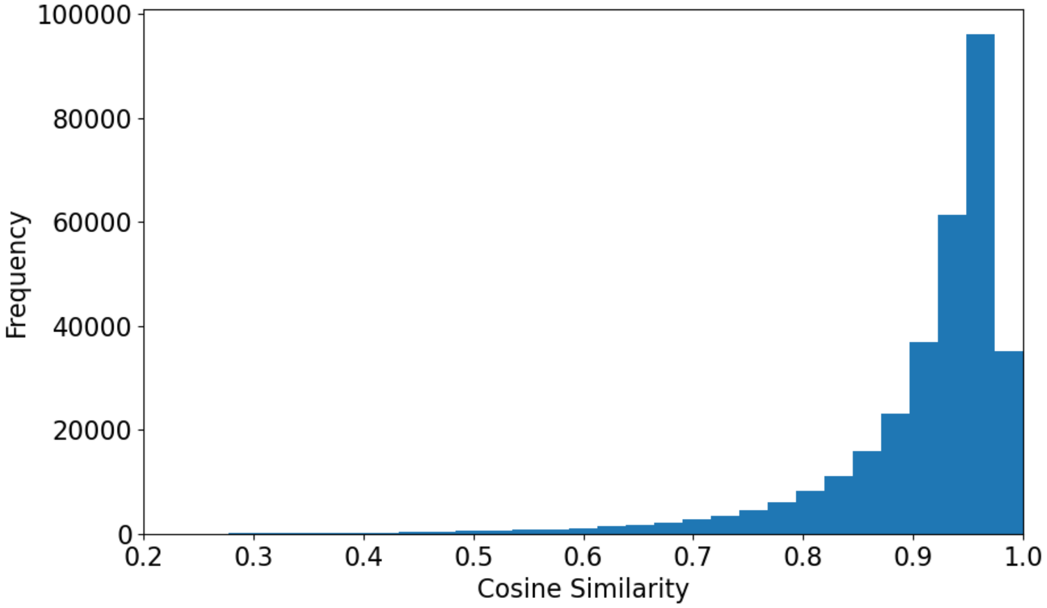

Figure A.7: Distribution of semantic similarities of all computer science papers, last five years

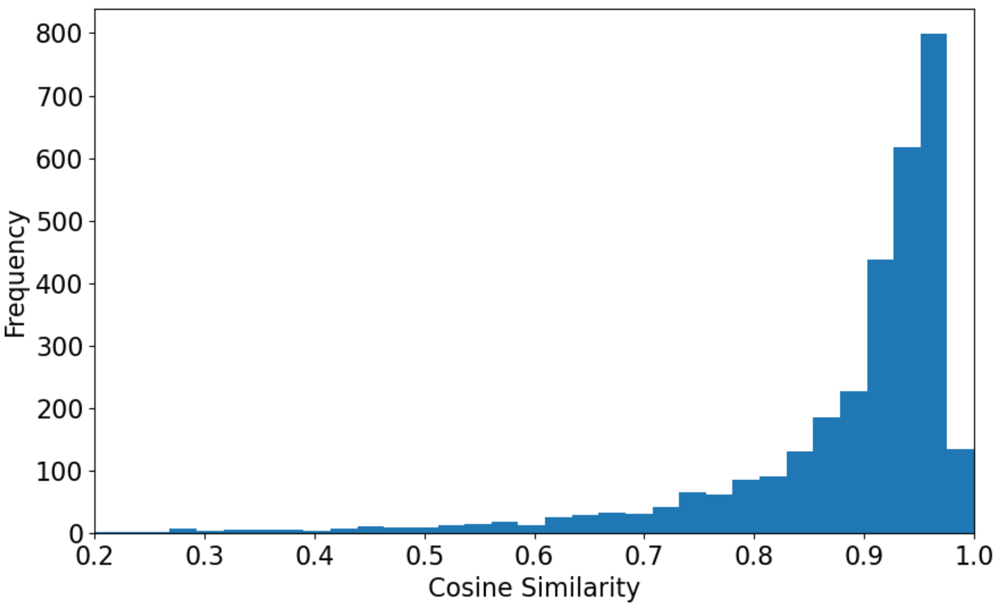

Figure A.8: Distribution of semantic similarities of all computer science papers, 2000-2004

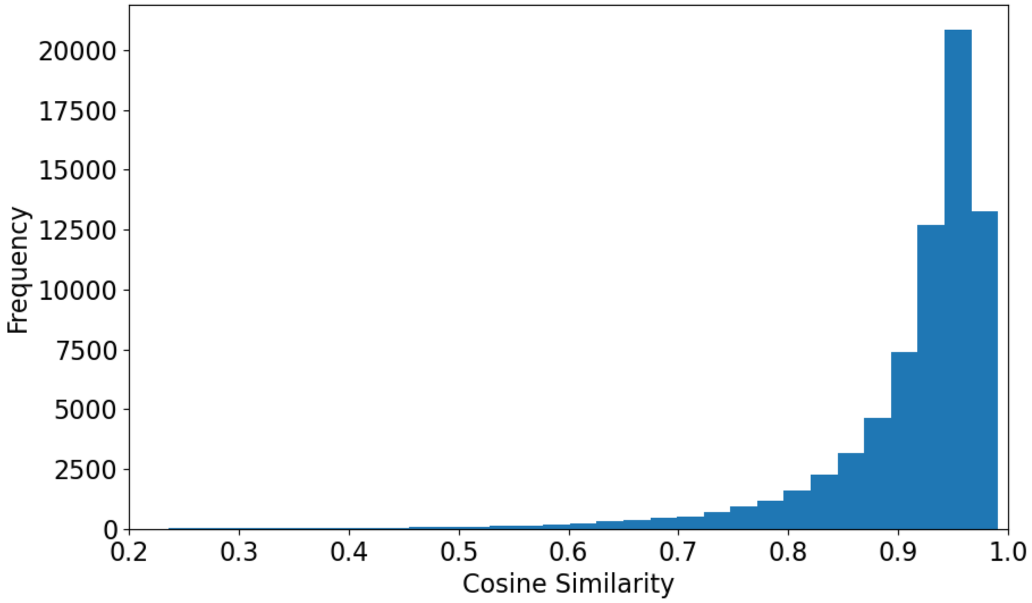

Figure A.9: Distribution of semantic similarities of all physics papers, last five years

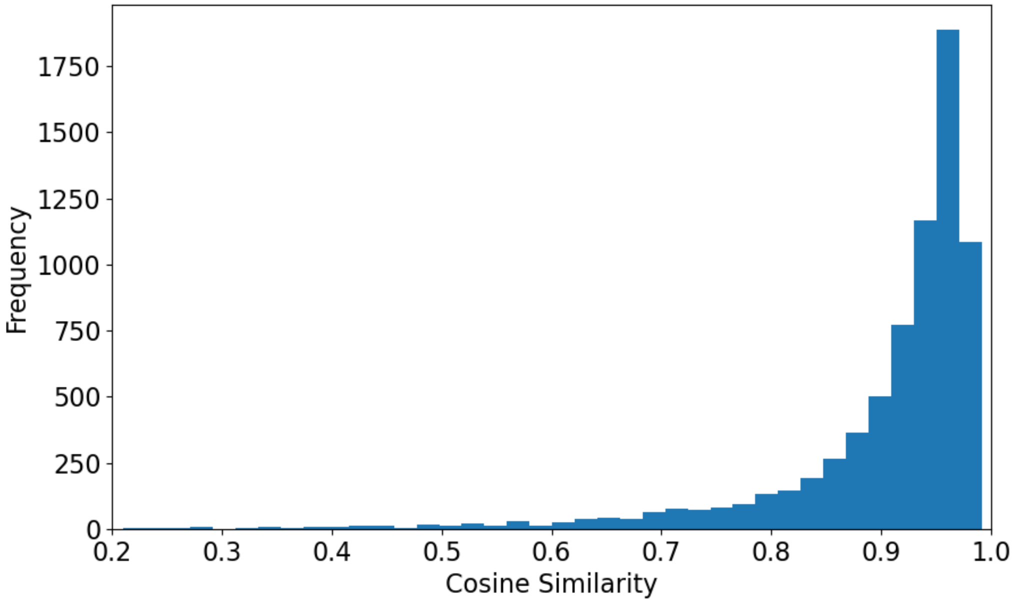

Figure A.10: Distribution of semantic similarities of all physics papers, 2000-2004

Mathematics	
Abbreviation	Subfield Name
math.CO	Combinatorics
math.CA	Classical Analysis
math.NT	Number Theory
math.PR	Probability Theory
math.NA	Numerical Analysis
math.RA	Rings and Algebras
math.OA	Operator Algebras
math.QA	Quantum Algebra
math.DG	Differential Geometry
math.FA	Functional Analysis
math.AG	Algebraic Geometry
math.DS	Dynamical Systems
math.GR	Group Theory
math.AC	Commutative Algebra
math.SG	Symplectic Geometry
math.GT	Geometric Topology
math.CV	Complex Variables
math.AP	Analysis of PDEs
math.RT	Representation Theory
math.MG	Metric Geometry
math.ST	Statistics Theory
math.AT	Algebraic Topology
math.OC	Optimization and Control
math.LO	Logic and Foundations

Computer Science	
Abbreviation	Subfield Name
cs.NE	Neural Networks
cs.DS	Data Structures
cs.CE	Computer Engineering
cs.IT	Information Theory
cs.CC	Computational Complexity
cs.DM	Discrete Mathematics
cs.CR	Cryptography
cs.NI	Networking and Internet Architecture
cs.PF	Performance and Reliability
cs.LG	Machine Learning
cs.CY	Cybersecurity
cs.CG	Computational Geometry
cs.CV	Computer Vision
cs.SE	Software Engineering
cs.OH	Other Subfields in Computer Science

cs.PL	Programming Languages
cs.AI	Artificial Intelligence
cs.IR	Information Retrieval
cs.GT	Game Theory
cs.LO	Logic in Computer Science
cs.SC	Symbolic Computation
cs.DC	Distributed Computing
cs.CL	Computational Linguistics
cs.HC	Human-Computer Interaction
cs.AR	Computer Architecture
cs.DL	Digital Libraries
cs.MS	Mathematical Software
cs.RO	Robotics
cs.DB	Databases
cs.GL	Computer Graphics
cs.MA	Multiagent Systems
cs.MM	Multimedia
cs.OS	Operating Systems
cs.NA	Numerical Analysis
cs.SD	Sound and Music Computing
cs.GR	Graphics
cs.FL	Formal Languages
cs.SI	Social Informatics
cs.SY	Embedded Systems
cs.ET	Emerging Technologies
Physics	
Abbreviation	Subfield Name
physics.gen-ph	General Physics
physics.optics	Optics
physics.ed-ph	Education in Physics
physics.pop-ph	Popular Physics
physics.soc-ph	Social Physics
physics.data-an	Data Analysis
physics.plasm-ph	Plasma Physics
physics.bio-ph	Biological Physics
physics.flu-dyn	Fluid Dynamics
physics.comp-ph	Computational Physics
physics.atom-ph	Atomic Physics
physics.chem-ph	Chemical Physics
physics.geo-ph	Geophysics
physics.class-ph	Classical Physics
physics.atm-clus	Atomic Clusters
physics.acc-ph	Accelerator Physics
physics.hist-ph	History of Physics
physics.ins-det	Instrumentation and Detectors

physics.space-ph	Space Physics
physics.med-ph	Medical Physics
physics.ao-ph	Atmospheric and Oceanic Physics
physics.app-ph	Applied Physics

Table A.1: Abbreviations and names of primary paper subfields